## FA 10.4: A 300MHz CMOS Microprocessor with Multi-MediaTechnology

Mustafiz R. Choudhury, James S. Miller

Intel Corp., Santa Clara, CA

This 7.5M transistor CMOS microprocessor is architecturally equivalent to, and the next generation superset of, a previous microprocessor with dynamic execution [1]. It implements Intel MMX™ Technology instructions to enhance performance of media and communication applications. It doubles the on-chip instruction and data caches of the previous generation processor to 16kB each and implements a dedicated bus to access large off-chip second-level caches (Figure 1). This processor also implements renaming of data segment registers. These and other features result in projected performance of ~12.0 Specint95 for this 300MHz processor and 3-12X performance improvement for kernels of multimedia and communication applications (for example DSP, video and voice) over a scalar implementation. It is the highest performance Intel-Architecture compatible microprocessor.

This processor implements 57 new MMX instructions that involve packed data types to support a single instruction multiple data (SIMD) technique whereby multiple data elements can be operated on in parallel [2]. A dual execution pipeline is used for these instructions. The arithmetic and logic operations hardware is duplicated in each of the pipelines but only one shifter and one multiplier are implemented to meet area constraints. They are placed in different pipelines so operations utilizing them can be executed in parallel (Figure 2). Implementation of MMX technology in this processor involved retrofitting the execution units in the microarchitecture of the previous-generation processor [1]. It required changes in instruction length decode, instruction decode, renamer, resource allocator and reservation station of the "[instruction fetch]-[length decode]-[instruction decode]-[rename and resource allocation]-[micro-op scheduling and dispatch]-[execution]-[writeback]-[retirement]" pipeline.

This processor also implements data segment register renaming to improve performance of segment loads. In the previous generation implementation, all segment register update operations are serialized resulting in performance impact. In this processor, each of the four data segment registers has their own single rename space. The implementation involves changes in the register alias table, the reservation station, address generation unit where copies of the registers are physically located and development of new microcode flows. Performance improvements of >8% over the previous generation are measured on Sysmark95 W3.11 at the same processor bus and core clock frequencies.

A dedicated second-level (L2) cache bus supports a variety of L2 cache sizes and configurations. These L2 caches can be constructed using synchronous burst static RAMs (BSRAMs) for data and optional ECC storage and the Intel 82459AB tagram for tag, state and LRU bits storage. Underlying the cache bus protocol is a resource tracking set of finite state machines and other control logic that support the wide range of cache sizes, latency, burst rates and data integrity features. The L2 cache bus employs specialized clocking achieved with a skew matching approach called source synchronous clocking. This enables a full-speed L2 cache interface for custom cache static RAMs (CSRAMs) as well as fractional speeds for BSRAM based L2 cache implementations.

This microprocessor adds new testability features for the large on-chip cache arrays. These features provide direct access to the large embedded memory arrays from the pins of the processor bus. It allows exhaustive testing of the arrays using automatic test pattern generator (ATPG) tester hardware without relying on the functionality of the microprocessor. Testability hooks to test the off chip L2 tag and storage arrays are implemented. Enhanced clocking allows system hardware to control most of the internal clocks and stop the external clock input to the PLL. The processor preserves the machine state and wakes up to a deterministic state when the external clock is applied.

This 203mm² microprocessor is implemented in a 0.35µm, 4-layer-metal CMOS process and packs in 7.5M transistors. This 2.8V process technology is optimized for logic products such as microprocessors in which chip performance is dictated by density, interconnect, clock frequency and power consumption [3]. The process provides aggressively-scaled CMOS devices with short channel lengths and four layers of Al-Cu planarized interconnect (Table 1). The top metal layer (metal-4) is thick, has low resistance, and is used for global power supply, clock, and critical signal routing. The lowest layer is used principally for local interconnect and the middle two for signals and local supplies.The clock generator in this microprocessor is programmable and can supply several integer and non-integer bus fraction (1/2,1/3,1/4,2/5,2/7,2/9) frequencies relative to the core frequency. It also generates the L2 cache bus clock that can run at an integer fraction (1, 1/2, 1/3) of the core frequency and maintain 50% duty cycle. Global clock is distributed to all units in metal 4. The global clock skew is <120ps, achieved by balancing the load of each global clock tapping and adjusting global clock track length. The power network is designed using the top 2-metal-layers (one vertical and one horizontal) to form a grid structure to minimize IR drop and AC noise, and meet electromigration requirements.

High-frequency design methodologies, tools, and circuit techniques are employed in the design. The performance and reliability of this chip relied heavily on strict adherence to well defined methodologies. This includes use of very accurate RC extract, timing analysis and reliability tools. Exhaustive parasitic extraction (both R and C) and static timing analysis of all circuits including cache memory circuits are employed to identify and fix frequency limiting critical paths. Domino circuits are widely used in the design to achieve high frequency. RC-dominated critical signals are routed in metal 4. A simple clocking scheme is used for dominos, latches and flip-flops that allow automated checks and compliance to methodology. Determination of cross coupling noise as a function of driver strength, line resistance, node capacitance, cross capacitance, and attacking line slope helps avoid functionality issues or timing degradation. Tools and methodology are used for electromigration and self-heating checks on all power and signal lines.

Special circuits are used in frequency-limiting blocks such as the SIMD adder and caches. The SIMD ALU adder (Figure 3) is capable of performing all variations (byte, word and long-word) of addition, subtraction with saturation (required to operate on packed data types defined in Intel MMX technology) and "compare greater than." In the data cache status bit array, a pMOS wired-or is used as the sense amplifier for bit lines that are finely split to reduce capacitance and increase performance. Progressive enabling is used for sense amplifiers of large cache arrays (Figure 4). The circuits in this processor are redesigned to eliminate bipolar transistors of the previous generation BiCMOS implementation. This redesign takes advantage of aggressively scaled CMOS devices and enables lower voltage power supplies.

The microprocessor described in this paper passes component level tests and correctly executes software (including those that utilize Intel MMX technology) in standard operating systems. It operates at 300MHz at 2.8V at 85°C.

0-7803-3721-2 / 97 / $5.00 / © IEEE

*Acknowledgments:*

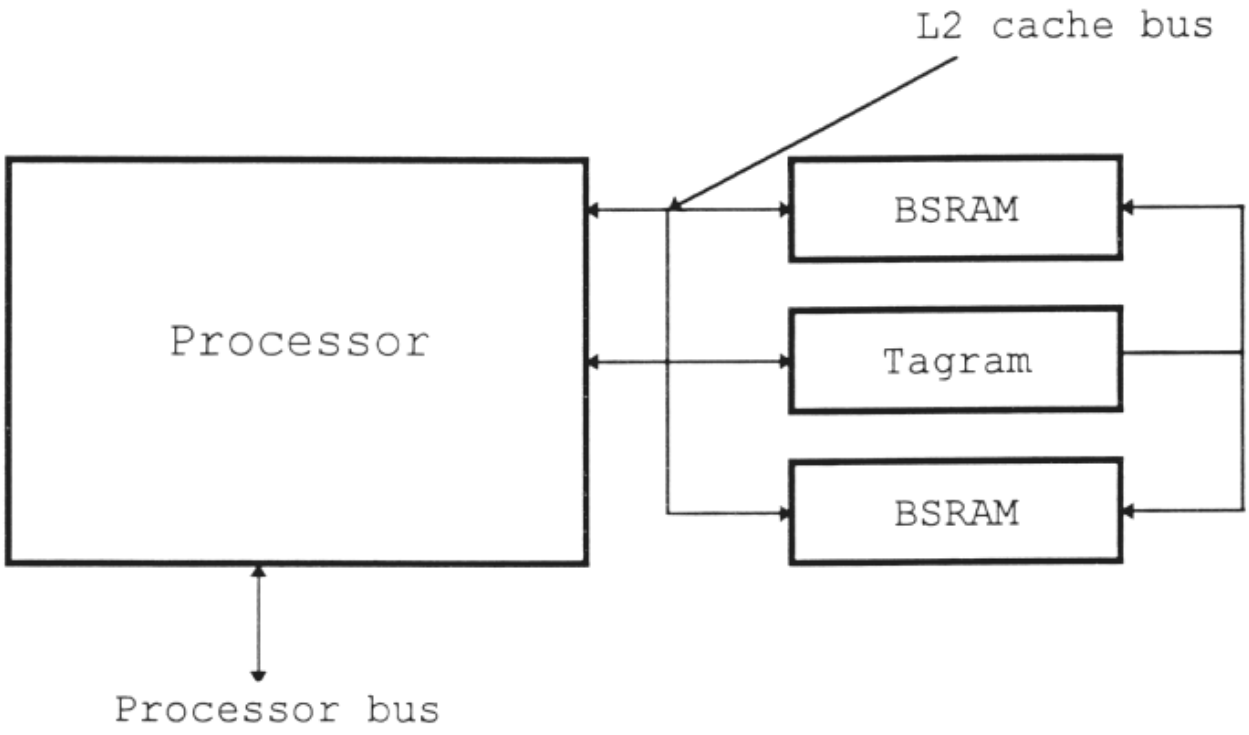
The work of a talented and dedicated team is presented in this paper. The authors feel privileged to represent their work.

*References:*

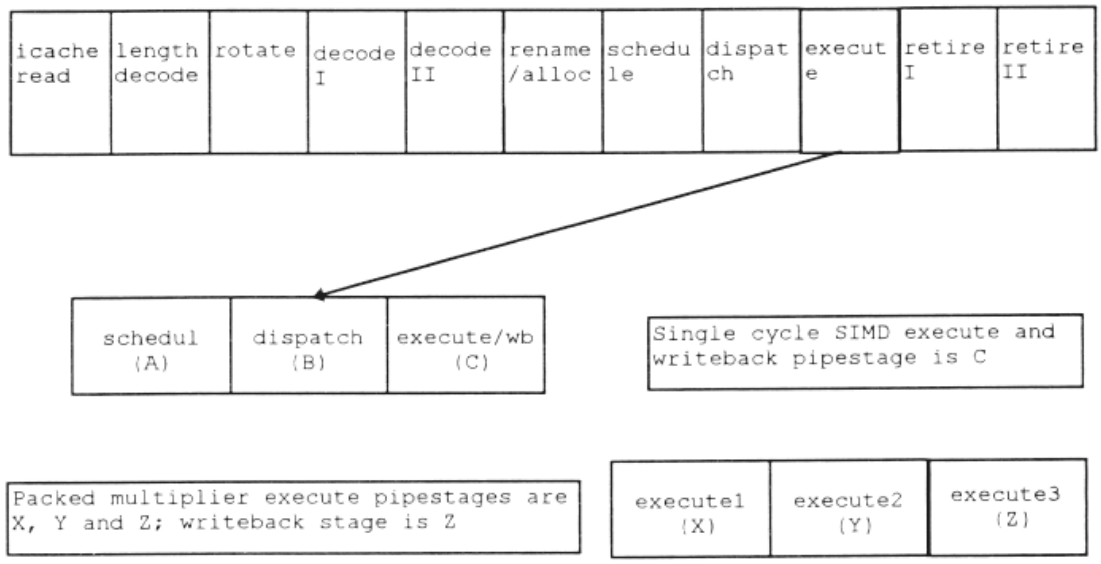
[1] Colwell, R., R. Steck, "A 0.6µm BiCMOS Processor with Dynamic Execution," ISSCC Digest of Technical Papers, pp. 176-177, Feb., 1995.

[2] Weiser, U., "Intel MMX Technology," Proceedings Hot Chips VIII Symposium, pp. 147-155, 1996.

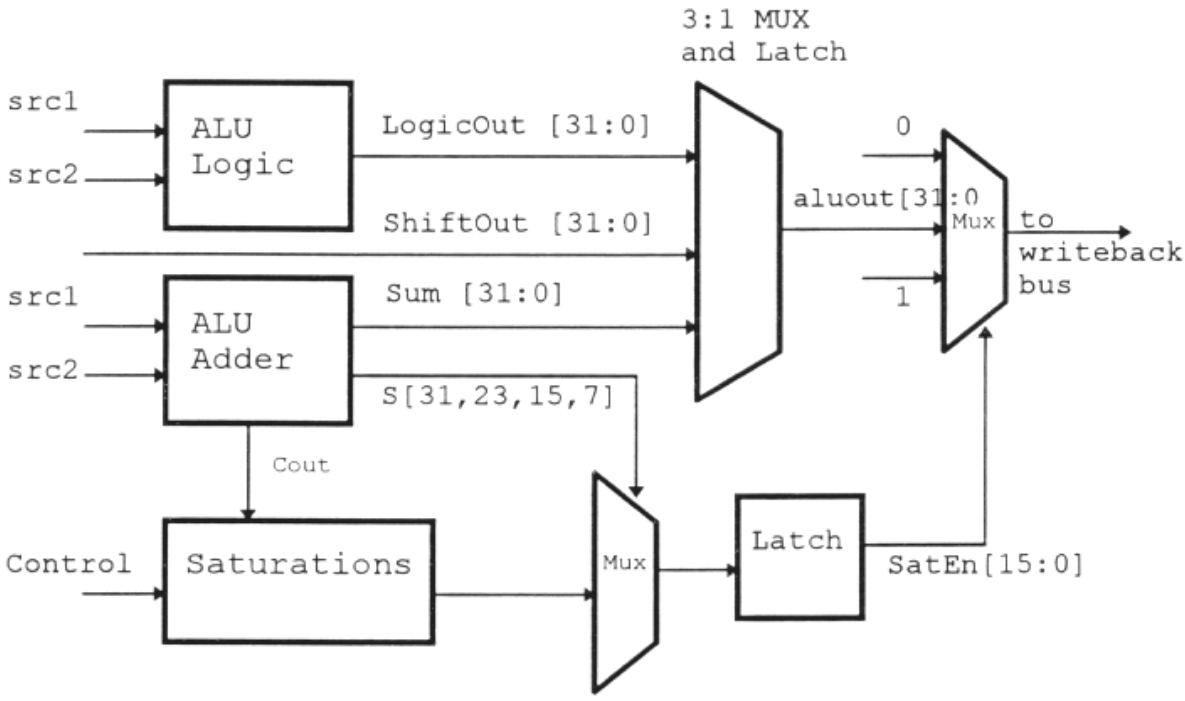
[3] Bohr, M., et al., "A High Performance 0.35µm Logic Technology for 3.3V and 2.5V operation," IEDM Technical Digest, pp. 273-276, 1994.



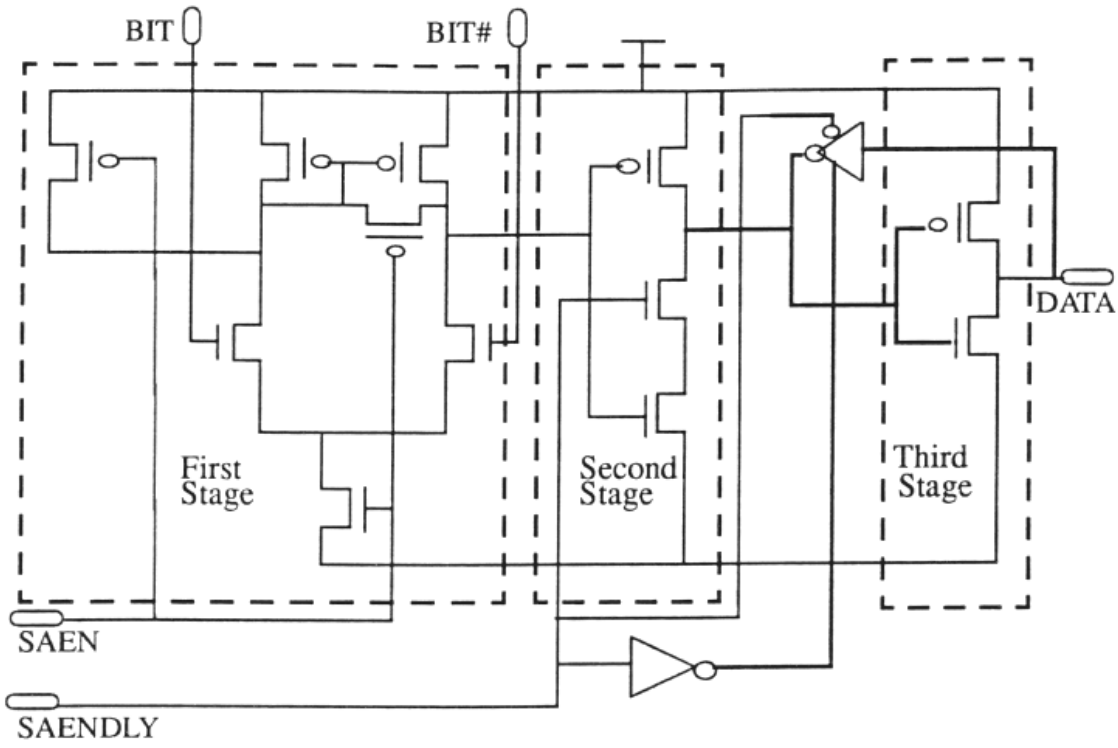

**10-4-1: Processor and off package L2 with Tagram and Burst SRAMs.**



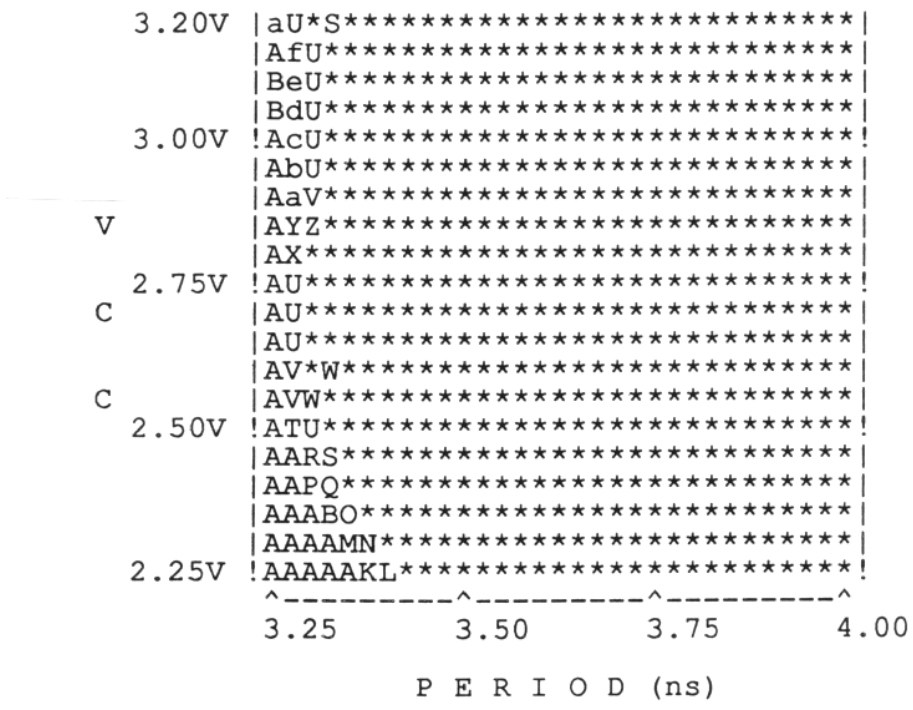

**10-4-2: SIMD execution unit pipeline stages.**




**10-4-3: SIMD ALU with saturation logic.**




**10-4-4: Progressive sense amp. enabling scheme for robust high-performance cache arrays.**




**10-4-5: Shmoo plot showing clock rate vs. supply voltage.**

| | |
|---|---|
| Die area | 203mm² |
| Supply voltage | 2.8V |
| Typical Leff | 0.22µm |
| Metal 1 pitch | 0.88µm |
| Metal 2 pitch | 1.16µm |
| Metal 3 pitch | 1.16µm |
| Metal 4 pitch | 3.04µm |
| Tox | 6nm |

**10-4-Table 1: Process details.**